DOCUMENT RESUME

ED 223 660 TM 820 763

AUTHOR Dole, Janice Arnold
TITLE The Convergent and Discriminant Validity of Literal and Inferential Comprehension Tests.
PUB DATE Apr 81
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981); Table 1 is marginally legible due to small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Rating; Grade 6; Intermediate Grades; Matrices; *Reading Comprehension; Reading Skills; *Reading Tests; *Standardized Tests; Teachers; *Test Validity
IDENTIFIERS *Inference Comprehension

ABSTRACT

The convergent and discriminant validities of tests designed to measure the subskills of literal and inferential reading comprehension in different content areas were assessed. A multitrait-multimethod matrix was generated, using three measuring methods: (1) a standardized reading test; (2) individually devised tests; and (3) teacher ratings. Subjects were 136 sixth-grade students. Results indicated moderate convergent validity and no discriminant validity for the subskills. Evidence from this study suggests that reading tests do not measure discrete skills in reading as well as many educators would believe, and that the reading process remains relatively stable across material in different content areas. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED223660

THE CONVERGENT AND DISCRIMINANT VALIDITY OF LITERAL AND INFERENTIAL COMPREHENSION TESTS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. A. Dole

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

by

Janice Arnold Dole, Ph.D.

Tm 820763

Paper presented to the
American Educational Research Association,
Los Angeles, California
April, 1981

ABSTRACT

This study assessed the convergent and discriminant validity of tests designed to measure the subskills of literal and inferential reading comprehension in different content areas. A multitrait-multimethod matrix was generated, using three measuring methods: (I) a standardized reading test; (2) individually devised tests; and (3) teacher ratings. Subjects were I36 sixth-grade students. Results indicated moderate convergent validity and no discriminant validity for the subskills. Evidence from this study suggests that reading tests do not measure discrete skills in reading as well as many educators would believe, and that the reading process remains relatively stable across material in different content areas.

The testing of reading comprehension is an issue of paramount importance to education, particularly as tests play a more important role in determining instructional and curriculum decisions and reading competency. Of particular concern to researchers is the validity of reading comprehension tests. Traditionally, these tests have measured several discrete and distinct subskills of reading. Whether these subskills are truly distinct, however, has been the subject of considerable debate (Andrich & Godfrey, 1978-79; Drahozal & Hanna, 1978; Farr, 1972; Flood & Lapp, 1978; Trela, 1967).

The purpose of this study was to assess the convergent and discriminant validity of tests designed to measure the subskills of literal and inferential comprehension. Can tests be designed to measure literal or inferential comprehension apart from each other and from a more global reading comprehension? A secondary purpose was to examine the validity of these subskills within different content areas. Is inferential comprehension a unique skill within one content area, or is it generalizable across all content areas?

For the purpose of this study literal reading was defined as the ability to answer questions about material directly stated in a passage. Inferential reading was defined as the ability to answer questions about material not directly stated in a passage.

The issue of separate skills of reading comprehension is not a new one. Several researchers have used factor analytic techniques in an attempt to isolate various subskills of reading comprehension (see, for example, Davis, 1944. 1968; Schreiner, Hieronymus & Forsyth, 1969; Spearritt, 1972). These researchers either used existing standardized tests or devised their own to measure the subskills (Drahozal & Hanna, 1978). Results of these and other efforts to isolate skills in reading have been and continue to be the subject of considerable debate. Some evidence suggests that reading can be separated into unique components, especially at the secondary level (Davis, 1944, 1968). Further, some researchers (see, for example, Singer, 1970) argue that reading may consist of a more generalized skill at the elementary level and more specialized skills at the secondary level. Nevertheless, results of studies reporting to isolate separate skills have been criticized (Spearritt, 1972; Thurstone, 1946), and contradicted by research which suggests that reading can be measured only as a global skill (Derrick, 1953; McCullough, 1957). After an extensive review of the literature on the measurement of reading skills, Lennon (1962, p. 333) concluded that "it seems entirely clear that numerous superficially discrete reading skills to which separate names or titles have been attached are, in fact, so closely related, as far as any test results reveal, that we must consider them virtually identical."

Literal and inferential comprehension are two of the many subskills identified as being discrete. Intuitively, it seems obvious that these skills are separate and distinct components of reading comprehension. Studies utilizing tests designed to measure the skills, however, have only had partial success in differentiating them empirically (Follman & Lowe, 1972-73; Maney, 1958; Sochor, 1958). The fact that many researchers reported high correlations between the two skills (Follman & Lowe, 1972-73; McCullough, 1957; Sullivan, 1975) indicates that items on the tests may be measuring the same skills rather than unique ones. In addition, the high correlations reported between either skill and reading in general indicate that these skills may not be measured apart from general reading ability.

Reading skills within content areas present a similar problem. Peters, Peters and Kaufman (1973) presented empirical data which support the intuitive notion that literal and inferential comprehension differ across content areas. They constructed content area reading tests and administered these tests along with a test of general reading ability to eighth grade students. Correlations between content area subtests of literal comprehension and reading were reported from .43 to .68. Peters et al. concluded that content area subskills were not as highly correlated with general reading ability as earlier research had indicated, and that results from one content area test were not generalizable to other content area

tests. Little empirical research exists which examines the relationship between these different measures of reading performance (Peters et al., 1973).

Method

Convergent and Discriminant Validity

The present study used a unique validational procedure, the multitrait-multimethod approach (Campbell & Fiske, 1959), to assess the validity of the subskills of literal and inferential comprehension in content area reading tests. Basically, in the multitrait-multimethod procedure, unique traits are correlated with unique methods of measuring those traits. Campbell and Fiske used the term "traits" to refer to psychological constructs, e.g., self-concept or impulsivity, but the "traits" also can refer to processes or skills. They used the term "methods" to refer to the different types of instruments, e.g., self-rating scales, paper and pencil tests, which measure the constructs. Convergent validity refers to the fact that different tests purporting to measure the same skills should correlate highly with each other. Discriminant validity refers to the fact that tests measuring different skills should demonstrate low correlations.

In the present study, literal and inferential comprehension and the content areas of science, social studies and English literature were the "traits" or skills to be measured. Each skill was crossed with literal and inferential comprehension to generate six "traits" or subskills: literal reading

in science, inferential reading in science, with similar components in social studies and English.

Instrumentation

In order to examine the convergent and discriminant validity of traits in the multitrait-multimethod matrix, the methods for measuring the traits need to be as distinct and independent as possible. This requirement presents a major dilemma in examining literal and inferential comprehension, as there are few unique methods of measuring those constructs. Traditionally, the measurement of reading comprehension has been limited to paper and pencil multiple-choice tests. Open-ended tests present major problems in interpretation. Tests administered orally are extremely time-consuming and add the additional factor of listening into the measurement procedure. The cloze procedure has not been shown to parcel out sub-skills and therefore could not be utilized as a unique measure of the constructs.

This study therefore utilized two different sets of paper and pencil tests as separate measures of the constructs. The tests differed in general purpose and intent, number of questions per passage, number of multiple-choice detractors per item, total length and in the utilization of speed as a function of comprehension. In addition, a more unique measurement, that of teacher ratings of students' abilities in literal and inferential comprehension, was utilized as a third measuring method.

The Intermediate Reading Tests: Social Studies and Science by Maney and Sochor (1958) were included in the matrix because both tests were specifically designed to measure literal and inferential comprehension. (Sochor and Maney used the terms "critial reading" for what is defined here as "inferential comprehension." Some reading educators define the terms broadly and use them synonymously; others differentiate the two as unique subskills in reading. For the purposes of this study, the terms will be used synonymously.) In order to complete the multitrait-multimethod matrix for this measuring method, it was necessary to construct a parallel test in the subject area of English literature. The Intermediate Reading Test: English was constructed utilizing the definitions and format of the Sochor and Maney tests.

The Sochor, Maney and Dole tests each consisted of three 250-300 word passages followed by approximately 22 items/66 items per test. Each test was untimed and utilized a five-item multiple choice format. The reliabilities for these tests, using Hoyt's internal consistency formula, were .82, .81, and .83, respectively.

A standardized reading test, the Comprehensive Test of Basic Skills (CTBS) was chosen as a second method for measuring the six subskills. This test was a timed test and consisted of five 250-300 word passages followed by 7-11 multiple-choice items per passage. Overall reliability for this test was

calculated using Hoyt's internal consistency formula as .80. The CTBS was designed as a global measure of reading comprehension rather than a specific measure of literal and inferential reading. Its inclusion into the matrix was predicated upon the fact that the CTBS is one of the most widely used and recognized readirg tests, and that administrators and teachers use the test to evaluate the reading sub-skills of their students. In addition, the CTBS measured all six sub-skills through a format different from the Sochor, Maney and Dole tests.

In order to utilize the information drawn from the CTBS in the multitrait-multimethod matrix, the test items were regrouped according to subject matter (content area) and level of comprehension (literal or inferential). Each passage and item was examined and classified into one of the six subskills according to the operational definitions of literal and inferential reading used in this study. For example,

Where did Mrs. Bethune open a hospital?

1. Florida
2. Georgia
3. South Carolina
4. Washington, D.C.

was classified as a literal comprehension item, and the item

Which of the following is the best title for the story?

1. "My Favorite Door"

2. "Door to the Future"

3. "Superstitious Doors"

4. "The Ins and Outs of Doors"

was classified as an inferential comprehension item. In this way new subtests were formed from the CTBS.

A third method, the Student Evaluation Scale, utilized subjective measures of students' abilities to read at the literal and inferential levels. Teachers of subjects in the study were given an evaluation scale to complete. Literal and inferential reading comprehension were defined for the teachers, and examples of both types of comprehension were included.

Teachers were asked to make subjective evaluations of their students' understanding at a literal and inferential level of comprehension in science, social studies and English. For example, the question was asked:

> How much information do you think this student can obtain at a literal level in science?

Teachers responded by circling one of four categories: "all," "most," "some" or "none."

Procedure

Subjects for the study were drawn from a middle to upper-middle class community in a western metropolitan area. The Sochor, Maney and Dole tests were administered on different days to 136 sixth graders from four schools within the same district. Results from the CTBS, administered three months

earlier, were made available to the researcher. The Student Evaluation Scale was given to all of the subjects' teachers and used as the third source of data.

Results

Content Area Matrix

Correlations were computed among all subskills and a matrix was constructed from the intercorrelations of all methods and measures. Reliabilities were calculated on all the basic data, and corrections were made for attenuation, thereby providing the relationships between "pure" traits or subskills. The multitrait-multimethod matrix, corrected for attenuation is presented in Table 1.

Insert Table 1 about here

Campbell and Fiske have identified four criteria for establishing convergent and discriminant validity using the multitrait-multimethod matrix. To establish convergent validity, the correlations in the validity diagonals (the underlined numbers in Table 1: e.g., .93, .82. .32, .87, .90, etc.) should be sizable and significantly greater than zero. A visual inspection of the matrix reveals moderate convergent validity for the six subskills. Thus, the measures in this matrix appear to be measuring common traits or subskills. The clear exception is found in literal reading in

social studies, where the validity diagonals are comparatively low, .32, .57, and .25.

To establish discriminant validity, Campbell and Fiske set three criteria for inspection of the multitrait-multimethod matrix. First, different instruments measuring the same subskills should correlate more highly than different instruments measuring different subskills. For example, English literal comprehension as measured by content area tests and the CTBS should correlate more highly than English literal comprehension measured by content area tests and inferential comprehension measured by the CTBS. This requirement is examined through a comparison of the underlined validity values (e.g., .93) with the corresponding values in the same row (.87, .99, .96, .82,.82) and column (.85, .49, .90, .72, .46) within the heteromethod (dotted) blocks. Evidence for discriminant validity exists when the correlation in the underlined validity value exceeds all corresponding correlations in the same row and column. Inspection of the six heteromethod blocks reveals no discriminant validity for any of the subskills. All of the underlined validity values are exceeded by corresponding heteromethod correlations in at least one or more rows and columns.

The second requirement for discriminant validity is met when measures of a given skill employing different methods correlate more highly than measures of different skills employing the same method. For example, measures of English literal

comprehension should correlate more highly with each other when measured by content area tests and the CTBS than should measures of English literal and inferential comprehension when measured by content area tests alone. This requirement involves a comparison of the underlined validity values (e.g., .93) with the correlations in the same row (.87, .93, .90, .80, 1.00) or column in the monomethod (solid line) blocks. In examining each of the underlined validity values in the matrix, lower corresponding correlations should be found in all of the monomethod blocks. Here again, no evidence for discriminant validity was found.

The third and last criteria for discriminant validity requires an inspection of the patterns of correlations observed in both the heteromethod and monomethod blocks. The same patterns of correlations should be observed in each of the blocks, e.g., monomethod values in the CTBS block, English literal--.96, .37, .95, .80, .68--should have a similar pattern to the heteromethod column below it, in this case .75, .63, .75, .75, .72, .66. An examination of the matrix again reveals no evidence for discriminant validity met through this final criteria.

Combined Content Area Matrix

An additional matrix was generated in which the two subskills of literal and inferential comprehension were measured by the three methods, content area tests, the CTBS and teacher ratings. In this matrix the three content areas

were combined so that content area was not added in as an additional factor in the matrix. The resulting combined content area matrix, corrected for attenuation, is presented in Table 2.

Insert Table 2 about here

The combined content area matrix examines the convergent and discriminant validity of literal and inferential comprehension measures by three methods. Examination of the validity values in the matrix according to the Campbell and Fiske criteria reveals moderate to strong evidence for convergent validity. In the critical area of discriminant validity, however, evidence is conflicting or lacking. The validity values for literal comprehension meet the second and third requirements of heteromethod and monomethod discrimination in two of three cases. Thus, some limited evidence for the validity of literal comprehension has been established through these last findings. On the other hand, no discriminant validity can be found for inferential comprehension; all of the correlations in both the heteromethod and monomethod blocks exceed those in the validity values.

Discussion

Both matrixes present moderate to strong evidence for convergent validity for the six subskills and for literal and

inferential comprehension. This apparent evidence for convergent validity makes more sense empirically in that moderate to high correlations can be expected when measuring subskills that might be closely related or based on underlying constructs such as general reading ability and intelligence. Reading subskills have typically shown a high correlation with each other (.60-.87), with general reading ability (.41-.79) and with intelligence (.48-.70) (Dole, 1977). Evidence here supports these high correlations and suggests that the tests may be measuring similar skills or abilities. It is clear also that literal and inferential comprehension in different content areas requires overlapping skills and abilities, at least as far as tests in this study indicate.

An alternative or additional explanation for the higher convergent validity values needs to be addressed. Subjects for the study were drawn from a middle to upper-middle class community and generally performed well on all tests. The negative skewness of this sample has the effect of systemically raising all of the correlations in the matrix. While Campbell and Fiske (1959) argued that the patterns of correlations do not change as a result of the skewness of the sample, the validity values do increase. Thus, the higher validity values could reflect the nonrepresentativeness of the sample rather than evidence for convergence of measures.

In the content area matrix, no evidence was found for the discriminant validity of the subskills. Thus, literal and inferential comprehension within each of the content areas could not be isolated as distinct subskills or traits. Campbell and Fiske posited two potential sources of error when a lack of discriminant validity exists: 1) the methods employed to measure the constructs (subskills) are more important in determining the final scores than are the constructs themselves; 2) the constructs (subskills) are inadequately operationalized or all too closely related to one another to be separate constructs.

This first source of error is evident when the matrix is examined carefully. Method-specific variance is indicated in that all of the monomethod blocks have consistently high correlations. This is particularly true for the teacher ratings, and generally reflects teachers' inability to differentiate the subskills. Campbell and Fiske contended that rating instruments are particularly prone to this lack of differentiation in that people tend to judge others in an overall manner. This argument can be substantiated through a visual inspection of the raw data on the Student Evaluation Scale. Generally, teachers tended to regard the subskills as similar or the same, thereby failing to differentiate to any degree. Students were therefore rated in a consistent manner.

Campbell and Fiske's second source of error, that the subskills are too closely related or inadequately operationalized, is probably reflected in the data as well. It has been argued by many that we do not have adequate measures of reading comprehension. This study utilized the most common methods of measuring comprehension, and these methods have failed to discriminate between particular subskills which have been intuitively and empirically identified as being unique. While test developers often include information about the convergent validity of their reading tests, they rarely report evidence for discriminant validity (Farr, 1972). Despite this fact, reading is most frequently taught as isolated and separate skills at the elementary level and as separate skills within content areas at the middle and secondary level. If it is difficult to isolate literal and inferential comprehension, then it is unlikely that a more refined delineation of skills would produce results substantially different from those in this study, since a review of the literature suggests that other reading skills are more closely related than are literal and inferential comprehension.

Conclusions

Construct validation is an ongoing process, and results of a single study can never be regarded as "proof" of the validity of a construct or a test. Results of this study do provide additional evidence that, at least in the upper elementary grades, reading tests do not measure discrete skills

in reading as well as reading teachers and administrators would believe. Tests that purport to measure reading comprehension may be better measures of a global reading ability than they are of discrete subskills.

In addition, data from this study support Carlson's (1975, p. 55) argument: "Previous writings on reading in the content areas have assumed that readers need to shift their modus operandi when reading different subject matter. These findings [Carlson's] suggest that the reading process remains relatively stable across passages from various content areas." Tests in this study did not reveal unique differences in the reading process across content areas. Differences in reading achievement in various content areas could reflect student motivation, interest and background knowledge rather than differences in the reading skills or processes themselves. The importance of these additional factors may have been underestimated in the past.

Table 1

Content Area Multitrait-Multimethod Matrix: Corrected for Attenuation

	Content Area Tests						CTBS Test						Teacher Ratings					
	ENGLR 1A	ENGINF 2A	SSLR 3A	SSINF 4A	SCLR 5A	SCINF 6A	ENGLR 1B	ENGINF 2B	SSLR 3B	SSINF 4B	SCLR 5B	SCINF 6B	ENGLR 1C	ENGINF 2C	SSLR 3C	SSINF 4C	SCLR 5C	SCINF 6C
1A																		
2A	1.00																	
3A	.80	.67																
4A	.90	.91	.88															
5A	.93	.84	.85	.83														
6A	.87	.89	.65	.88	.99													
1B	.93	.87	.99	.96	.82	.82												
2B	.85	.82	.69	.84	.62	.62	.96											
3B	.49	.64	.32	.43	.40	.50	.37	.30										
4B	.90	.88	.77	.87	.88	.99	.95	.95	.64									
5B	.72	.66	.90	.89	.90	.83	.80	.98	.40	.99								
6B	.46	.46	.64	.68	.58	.55	.68	.71	.28	.73	1.00							
1C	.57	.57	.57	.66	.55	.53	.75	.61	.29	.54	.68	.31						
2C	.54	.56	.53	.67	.49	.50	.63	.65	.22	.57	.59	.25	.82					
3C	.58	.62	.57	.68	.57	.59	.75	.65	.25	.59	.63	.27	.81	.79				
4C	.53	.57	.56	.63	.51	.51	.75	.63	.32	.56	.59	.32	.76	.77	.85			
5C	.59	.55	.53	.68	.55	.61	.72	.65	.19	.60	.70	.22	.79	.78	.91	.80		
6C	.51	.51	.52	.64	.46	.51	.66	.68	.25	.57	.68	.25	.73	.78	.78	.83	.81	

Note: The validity values are underlined.
The monomethod blocks are enclosed in solid lines.
The hetromethod blocks are enclosed in dotted lines.

Table 2

Combined Content Area Matrix: Corrected for Attenuation

		Content Area Tests		CTBS Tests		Teacher Ratings	
		Literal	Inferential	Literal	Inferential	Literal	Inferential
Content Tests	Literal						
	Inferential	.93					
CTBS	Literal	.97*	.96				
	Inferential	.86	.85	.76			
Teacher Ratings	Literal	.64	.70	.79*	.63		
	Inferential	.60	.65	.74	.65	.96	

*Asterisk indicates strongest convergent validity values and significant heteromethod, monomethod validity values.

References

Andrich, D., & Godfrey, J. R. Hierarchies in the skills of Davis' reading comprehension test, form D: An empirical investigation using a latent trait model. Reading Research Quarterly, 1978-79, 14(?), 182-200.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carlson, K. L. A different look at reading in the content areas. In W. D. Page (Ed.), Help for the reading teacher. Urbana: ERIC Clearinghouse on Research and Communication Skills, 1975.

Davis, F. B. Fundamental factors of comprehension in reading. Psychometrika, 1944, 9, 285-297.

Davis, F. B. Research in comprehension in reading. Reading Research Quarterly, 1968, 4, 499-545.

Derrick, G. Three aspects of reading comprehension as measured by tests of different lengths. Princeton, New Jersey: Educational Testing Service, RB-53-B, 1953.

Dole, J. A. A validation of the construct of critical reading and of three tests designed to measure critical reading ability (Doctoral dissertation, University of Colorado, 1977). Dissertation Abstracts International, 1978, 38, 7135A. (University Microfilms No. 5684)

Drahozal, E. C., & Hanna, G. S. Reading comprehension subscores: Pretty bottles for ordinary wine. Journal of Reading, 1978, 21(5), 416-420.

Farr, R. The convergent and discriminant validity of several upper level reading tests. In G. Schick (Ed.), Multidisciplinary aspects of college-adult reading, seventeenth yearbook of the National Reading Conference. Milwaukee, Wisconsin: The National Reading Conference, 1972.

Flood, J., & Lapp, D. Inferential comprehension: A grand illusion. Language Arts, 1978, 55(2), 188-191.

Follman, J., & Lowe, A. J. Empirical examination of critical reading and critical thinking--overview. Journal of Reading Behavior, 1972-73, 5, 159-168.

Lennon, R. T. What can be measured? The Reading Teacher, 1962, 15, 326-337.

Maney, E. S. Literal and critical reading in science. Journal of Experimental Education, 1958, 27, 57-64.

McCullough, C. M. Responses of elementary school children to common types of reading comprehension questions. Journal of Educational Research, 1957, 57, 65-70.

Peters, C. W., Peters, N. A., & Kaufman, B. D. A comparative analysis of reading comprehension in four content areas. In F. P. Greene (Ed.), Reading: The right to participate. Twentieth yearbook of the National Reading Conference. Milwaukee, Wisconsin: The National Reading Conference, 1973.

Schreiner, R. L., Nieronymus, A. N., & Forsyth, R. Differential measurement of reading abilities at the elementary school level. Reading Research Quarterly, 1969, 5(1), 84-99.

Singer, H. Factors involved in general reading ability and reading in the content areas. In G. B. Schick & M. M. May (Eds.), Reading: Process and pedagogy, Vol. I. Nineteenth yearbook of the National Reading Conference. Milwaukee, Wisconsin: The National Reading Conference, 1970.

Sochor, E. E. Literal and critical reading in social studies. Journal of Experimental Education, 1958, 27, 49-56.

Spearritt, D. Identification of subskills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly, 1972, 8, 92-111.

Sullivan, J. The relationship of creative and convergent thinking to literal and critical reading ability of children in the upper grades. Journal of Educational Research, 1973, 66, 374-377.

Thurstone, L. L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

Trela, T. M. Comparing achievement of tests of general and critical reading. Journal of the Reading Specialist, 1967, 6, 140-142.